Internet Routing of Cloud Data by Using BGP

G. Komala

Asst. Professor, Computer Science & Engineering,
Christu Jyoti Institute of Technology & Science, Jangaon, Warangal, India

Abstract— Today's Internet often suffers transient outages, but as increasingly critical services migrate to the cloud; much higher levels of Internet availability will be necessary. The stunning shift toward cloud computing has created new pressures on the Internet. Loads are soaring, and many applications increasingly depend on real-time data streaming. Unfortunately, the reliability of Internet data streaming leaves much to be desired. Here, we focus on routing in the Internet's core, at extremely high data rates (all-to-all data rates of 40 Gbits per second are common today, with 100 Gbits/s within sight). These kinds of routers are typically implemented as clusters of computers and line cards; in effect a data center dedicated to network routing. The architecture is such that individual components can fail without bringing the whole operation to a halt. . For example, network links are redundant; if one link fails, there will usually be a backup. Such a router could even run routing protocols of different type's side by-side, making the actual routing decisions by consensus — if some protocol instance malfunctions, its peers would simply outvote it. But suppose that a routing protocol (for clarity, we focus on the Border Gateway Protocol [BGP], implemented by a BGP daemon [BGPD] hosted on some node within the router) needs to be restarted after a crash or updated with a software patch or migrated within the cluster.

Keywords: Cloud Computing, Network, Routing, clusters, Border Gateway Protocol.

I. Introduction

Cloud computing, particularly in conjunction with increased device mobility, is reshaping the Internet. We're seeing unprecedented shifts in demand patterns, a broad spectrum of new quality expectations, and a realignment of the entire field's economics. The implications are far-reaching.

The main text of this article focuses on high availability, one of several key properties today's cloud computing applications demand. The need is most obvious in voice-over-IP (VoIP) telephony and video disruptions can cause connections to seize up or fail in ways streaming: for such uses, even the briefest that are highly visible to the end user.

If we can crack the "high availability barrier," we can imagine a future in which the Internet carries all such traffic. Many cloud computing uses are so important (both in the terms of their scale and the associated revenue streams) that unless the Internet can evolve to meet the demands, the associated cloud computing enterprises might consider building new networks that would be dedicated to their use. For example, network links are redundant; if one link fails there will usually be a backup. Such a router could even run routing protocols of different types side -by-side, making the actual routing decisions by consensus- if some protocol instance malfunctions, its peers would simply outvote it. But suppose that a routing protocol (for clarity, we focus on the Border Gateway Protocol [BGP], implemented by a BGP daemon [BGPD] hosted on some node within the router) needs to be restarted after a crash or updated with a software patch or migrated within the cluster.

A Close Look at BGP

BGP is a very robust and scalable routing protocol, as evidenced by the fact that BGP is the routing protocol employed on the Internet. At the time of this writing, the Internet BGP routing tables number more than

90,000 routes. To achieve scalability at this level, BGP uses many route parameters, called attributes, to define routing policies and maintain a stable routing environment.

In addition to BGP attributes, classless inter domain routing (CIDR) is used by BGP to reduce the size of the Internet routing tables. For example, assume that an ISP owns the IP address block 195.10.x.x from the traditional Class C address space. This block consists of 256 Class C address blocks, 195.10.0.x through 195.10.255.x. Assume that the ISP assigns a Class C block to each of its customers. Without CIDR, the ISP would advertise 256 Class C address blocks to its BGP peers. With CIDR, BGP can supernet the address space and advertise one block, 195.10.x.x. This block is the same size as a traditional Class B address block. The class distinctions are rendered obsolete by CIDR, allowing a significant reduction in the BGP routing tables. BGP neighbors exchange full routing information when the TCP connection between neighbors is first established. When changes to the routing table are detected, the BGP routers send to their neighbors only those routes that have changed. BGP routers do not send periodic routing updates, and BGP routing updates advertise only the optimal path to a destination network.

II. Internal and External BGP

A BGP router can communicate with other BGP routers in its own AS or in other ASs. Both the I-BGP and E-BGP implement the BGP protocol with a few different rules. All I-BGP-speaking routers within the same AS, must peer with each other in a fully connected mesh. They are not required to be physical neighbors, just to keep a TCP connection as a reliable transport mechanism. Because there is no loop detection mechanism in I-BGP, all I-BGP-speaking routers must not forward any 3rd-party routing information to their peers. In contrast, E-BGP routers are able to advertise 3rd-party information to their E-BGP peers, by default. Figure 1: shows routers R1, R2, and R3 using I-BGP to exchange routing information within the same AS, and router pairs R4-R2, R3-R5, and R4-R5 using E-BGP to exchange routing information between Ass.

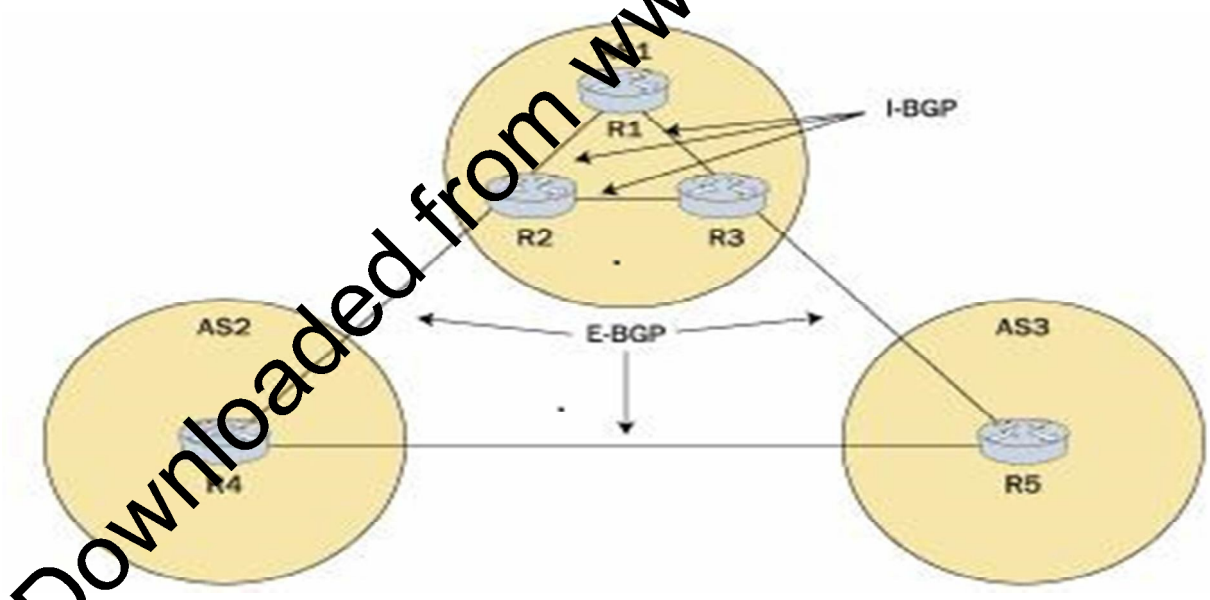

Figure 1.Internal BGP (I-BGP) versus external BGP (E-BGP).

BGP is designed for use in networks composed of interconnected autonomous systems (ASs). An AS could be a network operated by some ISP, or might be a campus or corporate rate network. BGP maintains a table of IP networks, or “prefixes,” that represent paths to a particular AS or set of ASs, tracking both direct neighbors and more remote ones. A BGPD instance runs on a router and uses path availability, network policies, or operator-defined databases of routing rules to select preferred routes.

III. Operation

BGP neighbors, called peers, are established by anual configuration between routers to create a TCP session on port 179. A BGP speaker sends 19-byte keep-alive messages every 30 seconds to maintain the connection.[3] Among routing protocols, BGP is unique in using TCP as its transport protocol. When BGP runs between two peers in the same autonomous system (AS), it is referred to as Internal BGP (iBGP or Interior Border Gateway Protocol). When it runs between different autonomous systems, it is called External BGP (EBGP or Exterior Border Gateway Protocol). Routers on the boundary of one AS exchanging information with another AS are called border or edge routers or simply eBGP peers and are typically connected directly, while iBGP peers can be interconnected through other intermediate routers. Other deployment topologies are also possible, such as running eBGP peering inside a VPN tunnel, allowing two remote sites to exchange routing information in a secure and isolated manner. The main difference between iBGP and eBGP peering is in the way routes that were received from one peer are propagated to other peers. For instance, new routes learned from an eBGP peer are typically redistributed to all other iBGP peers as well as all eBGP peers (if transit mode is enabled on the router). However, if new routes were learned on an iBGP peering, then they are re-advertised only to all other eBGP peers. These route-propagation rules effectively require that all iBGP peers inside an AS are interconnected in a full mesh.

Filtering routes learned from peers, their transformation before redistribution to peers or before plumbing them into the routing table is typically controlled via route-maps mechanism. These are basically rules which allow to apply certain actions to routes matching certain criteria on either ingress or egress path. These rules can specify that the route is to be dropped or, alternatively, its attributes are to be modified. It is usually the responsibility of the AS administrator to provide the desired route-map configuration on a router supporting BGP. Finite-state machines In order to make decisions in its operations with peers, a BGP peer uses a simple finite state machine (FSM) that consists of six states: Idle; Connect; Active; OpenSent; OpenConfirm; and Established. For each peer-to-peer session, a BGP implementation maintains a state variable that tracks which of these six states the session is in.

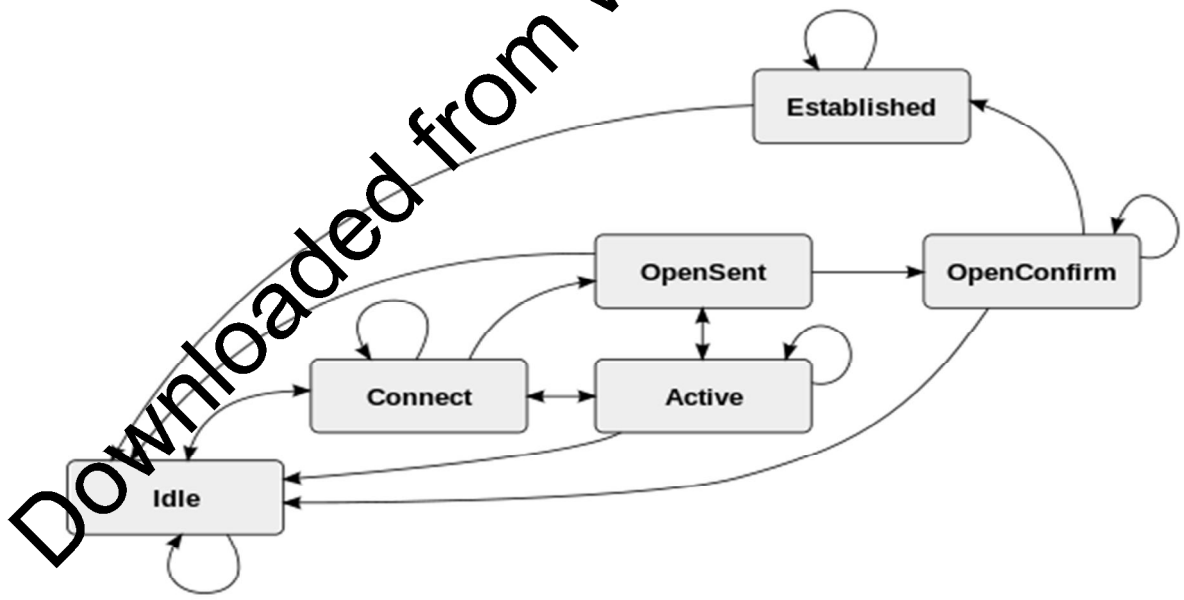

Figure 2. BGP state machine

The BGP defines the messages that each peer should exchange in order to change the session from one state to another. The first state is the "Idle" state. In the "Idle" state, BGP initializes all resources, refuses all inbound BGP connection attempts and initiates a TCP connection to the peer. The second state is

"Connect". In the "Connect" state, the router waits for the TCP connection to complete and transitions to the "Open Sent" state if successful. If unsuccessful, it starts the Connect Retry timer and transitions to the "Active" state upon expiration. In the "Active" state, the router resets the Connect Retry timer to zero and returns to the "Connect" state. In the "Open Sent" state, the router sends an Open message and waits for one in return in order to transition to the "Open Confirm" state. Keep alive messages are exchanged and, upon successful receipt, the router is placed into the "Established" state. In the "Established" state, the router can send/receive: Keep alive; Update; and Notification messages to/from its peer.

Idle State

- Refuse all incoming BGP connections
- Start the initialization of event triggers.
- Initiates a TCP connection with its configured BGP peer. Listens for a TCP connection from its peer.
- Changes its state to Connect.
- If an error occurs at any state of the FSM process, the BGP session is terminated
- Immediately and returned to the Idle state. Some of the reasons why a router does not progress from the Idle state are:
- TCP port 179 is not open.
- A random TCP port over 1023 is not open.
- Peer address configured incorrectly on either router.
- As number configured incorrectly on either router.

Connect State

- Waits for successful TCP negotiation with peer.
- BGP does not spend much time in this state if the TCP session has been successfully established.
- Sends Open message to peer and changes state to Open Sent.
- If an error occurs, BGP moves to the Active state. Some reasons for the error are:
- TCP port 179 is not open.
- A random TCP port over 1023 is not open.
- Peer address configured incorrectly on either router.
- AS number configured incorrectly on either router.

Active State

- If the router was unable to establish a successful TCP session, then it ends up in the Active state.
- Repeated failures may result in a router cycling between the Idle and Active states. Some of the reasons for this include:
- TCP port 179 is not open.
- A random TCP port over 1023 is not open.
- BGP configuration error.
- Network congestion.
- Flapping network interface.

Open Sent State

- BGP FSM listens for an Open message from its peer.
- Once the message has been received, the router checks the validity of the Open message.
- If there is an error it is because one of the fields in the Open message does not match between the peers, e.g., BGP version mismatch, MD5 password mismatch, the peering router expects a different My AS, etc. The router then sends a Notification message to the peer indicating why the error occurred.

- If there is no error, a Keepalive message is sent, various timers are set and the state is changed to Open Confirm.

Open Confirm State

- The peer is listening for a Keep alive message from its peer.
- If a Keep alive message is received and no timer has expired before reception of the Keep alive, BGP transitions to the Established state.
- If a timer expires before a Keep alive message is received, or if an error condition occurs, the router transitions back to the Idle state.

Established State

- In this state, the peers send Update messages to exchange information about each route being advertised to the BGP peer.
- If there is any error in the Update message then a Notification message is sent to the peer, and BGP transitions back to the Idle state.
- If a timer expires before a Keep alive message is received, or if an error condition occurs, the router transitions back to the Idle state.

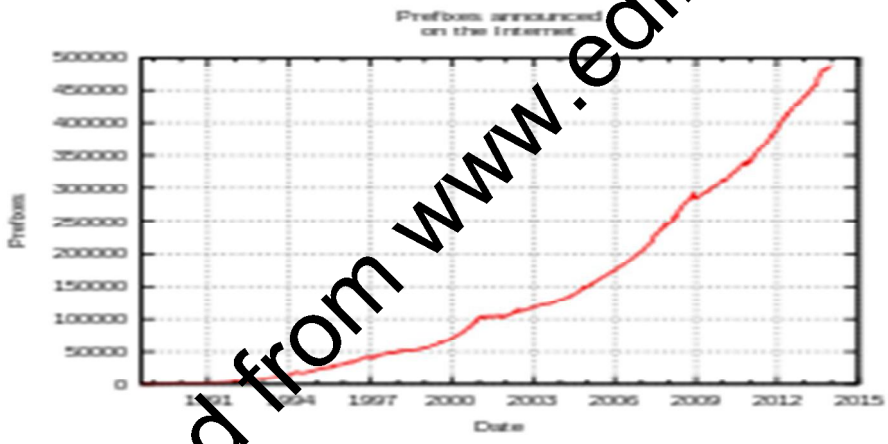

Figure 3. Growth on the Internet

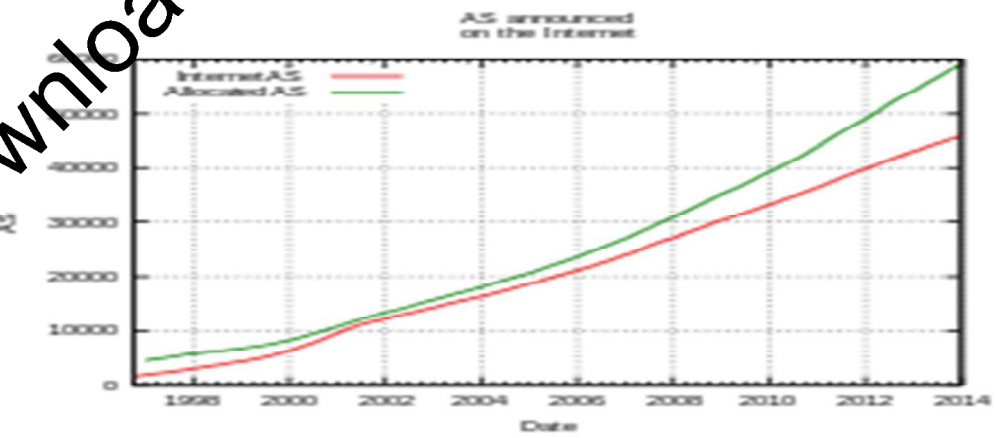

Figure 4. Number of AS on the Internet vs. number of registered AS.

One of the largest problems faced by BGP, and indeed the Internet infrastructure as a whole, is the growth of the Internet routing table. If the global routing table grows to the point where some older, less capable, routers cannot cope with the memory requirements or the CPU load of maintaining the table, these routers will cease to be effective gateways between the parts of the Internet they connect. In addition, and perhaps even more importantly, larger routing tables take longer to stabilize (see above) after a major connectivity change, leaving network service unreliable, or even unavailable, in the interim.

Until late 2001, the global routing table was growing exponentially, threatening an eventual widespread breakdown of connectivity. In an attempt to prevent this, ISPs cooperated in keeping the global routing table as small as possible, by using Classless Inter-Domain Routing (CIDR) and route aggregation. While this slowed the growth of the routing table to a linear process for several years, with the expanded demand for multi homing by end user networks the growth was once again super linear by the middle of 2004. A full IPv4 BGP table as of September 2012 is in excess of 430,000 prefixes.[17]

Route summarization is often used to improve aggregation of the BGP global routing table, thereby reducing the necessary table size in routers of an AS. Consider AS1 has been allocated the big address space of 172.16.0.0/16, this would be counted as one route in the table, but due to customer requirement or traffic engineering purposes, AS1 wants to announce smaller, more specific routes of 172.16.0.0/18, 172.16.64.0/18 and 172.16.128.0/18. The prefix 172.16.192.0/18 does not have any hosts so AS1 does not announce a specific route 172.16.192.0/18. This all counts as AS1 announcing four routes.

AS2 will see the 4 routes from AS1 (172.16.0.0/16, 172.16.0.0/18, 172.16.64.0/18 and 172.16.128.0/18) and it is up to the routing policy of AS2 to decide whether or not to take a copy of the four routes or, as 172.16.0.0/16 overlaps all the other specific routes, to just store the summary, 172.16.0.0/16.

If AS2 wants to send data to prefix 172.16.192.0/18, it will be sent to the routers of AS1 on route 172.16.0.0/16. At AS1's router, it will either be dropped or a destination unreachable ICMP message will be sent back, depending on the configuration of AS1's routers.

If AS1 later decides to drop the route 172.16.0.0/16, leaving 172.16.0.0/18, 172.16.64.0/18 and 172.16.128.0/18, AS1 will drop the number of routes it announces to three. AS2 will see the three routes, and depending on the routing policy of AS2, it will store a copy of the three routes, or aggregate the prefix's 172.16.0.0/18 and 172.16.64.0/18 to 172.16.0.0/17, thereby reducing the number of routes AS2 stores to only two: 172.16.0.0/17 and 172.16.128.0/18.

If AS2 wants to send data to prefix 172.16.192.0/18, it will be dropped or a destination unreachable ICMP message will be sent back at the routers of AS2 (not AS1 as before), because 172.16.192.0/18 would not be in the routing table.

IV .Future Work

Previous research used poisoning as a measurement tool to uncover hidden network topology and to assess the prevalence of default routes. While inspired by this work, ours differs in that we propose using poisoning operationally as a means to improve Internet availability.

Ongoing work seeks to verify the origin of BGP . By allowing an AS to poison only prefixes it originates, our approach is consistent with that goal. Proposals to verify the entire path are also consistent with our general approach, if we consider the poison as a (validated) hint from the origin AS to the rest of the network that a particular AS is not correctly routing its traffic. By the time such proposals are deployed, it should be feasible to develop new routing primitives or standardized BGP.

V. Conclusion

Cloud computing is the most network centric compute paradigm to date. A successful transition to cloud will depend on a rock solid network foundation .Today's cloud computing systems are appealing for their low cost of ownership, amazing scalability, and flexibility. The cloud even brings environmental benefits: users share computing resources, which are used more efficiently, and the data centers are typically located near power generating sources: by using the net generating sources: by using the network routing instabilities make the cloud less reliable than it needs to be.

VI. References

1. C. Laborite, G.R. Malan, and F. Jahanian, "Internet Routing Instability," IEEE/ACM Trans. Networking, vol. 6, no. 5, 1998, pp. 515–526.
2. E. Keller, J. Rexford, and J. van der Merwe, "Seamless BGP Migration with Router Grafting," Proc. Networked Systems Design and Implementation (NSDI 10), Use nix Assoc., 2010, pp. 16–30.
3. Orbit-Computer-Solutions.Com (n.d), Computer Training & CCNA Networking Solutions, Orbit-Computer-Solutions.com, retrieved 8 October 2013.
4. Network Working Group 2006, RFC 4271, Standards Track, retrieved 8 October 2013, <http://tools.ietf.org/html/rfc4271>.
5. Capabilities Advertisement with BGP-4, RFC 2842, R. Chandra & J. Scudder,May 2000
6. Multiprotocol Extensions for BGP-4, RFC 2858, T. Bates et al,June 2000
7. BGP/MPLS VPNs., RFC 2547, E. Rosen and Y. Rekhter,April 2004
8. IANA registry for BGP Extended Communities Types, IANA,2008
9. IETF drafts on BGP signaled QoS, Thomas Knoll,2008
10. BGP Route Reflection: An Alternative to Full Mesh Internal BGP (iBGP), RFC 4456, T. Bates et al., April 2006
11. Autonomous System Confederations for BGP, RFC 5065, P. Traina et al., February 2001